Comparison of Different Machine Learning Models for Cancer Prediction

Bhavya Sharma[1], Mohit Gambhir[2], Dr. Sapna Gambhir[3]
[1] *J.C. Bose University of Science and Technology YMCA, Faridabad, Haryana, India*
[2]*Verispire Technologies Pvt. Ltd., India*
[3] *J.C. Bose University of Science and Technology YMCA, Faridabad, Haryana, India*

(E-mail: bhavyasharma2164@gmail.com)

***Abstract*—** Malignant growth is the second driving reason for death all around and represented 9.1 million died in year 2018 itself. It has been portrayed as a heterogeneous ailment comprising of a wide range of subtypes. The early determination and guess of a malignant growth type have turned into a need in disease investigate, as it can encourage the resulting clinical administration of patients. For better clinical choices, it is critical to precisely recognize benevolent and harmful tumors. Expectedly, measurable strategies have been utilized for characterization of high hazard and generally safe malignancy, in spite of the perplexing cooperation's of high-dimensional restorative information. To defeat the downsides of customary factual strategies, machine learning has developed as a promising method for taking care of high-dimensional information, with expanding application in clinical choice help. This paper or scheme features new research bearings and talks about the fundamental difficulties identified with machine learning approaches in malignancy identification and forecasting**.**

***Keywords*—** *Machine Learning, Cancer Prediction and Diagnostic, Feature Selection, Artificial Neural Networks, Linear Regression, Support Vector Machine, Naïve Bayes.*

I. INTRODUCTION

Malignancy is definitely not a solitary ailment, but instead many related sicknesses that all include uncontrolled cell development and multiplication. It is driving reason for death in the created world and second in the creating scene, killing just about 9 million individuals every year. The early analysis and forecast of a disease type have turned into a need in malignant growth examine, as it can encourage the consequent clinical administration of patients and for better clinical choices, it is vital to precisely recognize considerate and harmful tumors. Expectedly, factual and computational techniques have been utilized for the recognition of high hazard and generally unsafe malignant growth, not withstanding the mind-boggling associations using linear regression and KNN of high dimensional therapeutic statistical information would be extracted for appropriate and accurate understanding of cancer conditions. To defeat the downsides of customary factual strategies, all the more as of machine leading has been connected to malignant growth visualization and expectation. Machine Learning is a part of man-made consciousness that utilizes an assortment of factual, probabilistic and improvement strategies that enables computational models and frameworks to "learn" from past models and to recognize hard to observe designs from expansive, loud or complex informational indexes. This ability is especially appropriate to medicinal applications, particularly those that rely upon complex proteomic and genomic estimations using feature extraction and confusion matrix. Accordingly, machine leading is as often as possible utilized in malignant growth analysis and identification. various methodologies are especially intriguing as it is a piece of a developing pattern towards customized, prescient medication and forecasting scenarios. Various patterns are utilized, including a developing reliance on protein biomarkers and microarray information, a solid inclination towards applications in cancer malignant growth, and a substantial dependence on "more seasoned" innovations such artificial neural networks (ANNs) and various techniques and facilitation created more effectively interpretable machine leading strategies. Various distributed examinations likewise seem to come up short on a suitable dimension of approval or testing. Among the better structured and approved investigations, obviously machine learning techniques can be utilized to considerably (15– 25%) improve the precision of anticipating disease helplessness, repeat and mortality. At an increasingly major dimension, it is likewise obvious that machine learning is additionally improving our fundamental comprehension of malignant growth advancement and movement. In this review distinctive model inculcated using linear regression and KNN of machine learning technique to elevate information for forecasting and diagnostic mechanism being coordinated and the execution of these strategies will produce the more effective and accurate forecast using various attributes, weights and measures.

Confront representation based on taxonomy of malignancy patients: The achievements in present days are to develop proof-based and customized restorative research is profoundly reliant on the accessibility of an adequate information premise as far as amount and quality. This frequently additionally suggests subjects like Neural Networks and AI for information retrieval. In the region of contention between information protection, institutional structures, and research interests, a few specialized, hierarchical and lawful difficulties arises to espionage the potential information in the field of cancer and its forecasting system. Adapting to these difficulties is one of the primary assignments of data the board in therapeutic research. In disease investigate, contextual investigation calls

attention to the peripheral conditions, necessities and quirks of taking care of research information with regards to therapeutic research. Seeing exploration results, it ends up clear that malignant growth sicknesses are increasingly similar to infection families with a huge number of sub-types and that the anatomical grouping of tumors may deceive and an arrangement as indicated by the neurotic difference in flagging pathways on the cell level is progressively satisfactory. The separation is significant on the grounds that for one patient a specific treatment might be viable and totally pertinent while it has no positive effect on tumor control for different patients with the "same" malignant growth and just bears reactions. So as to have a proof based medication with a sound measurable premise, the sum and nature of accessible information turn out to be essential. The required measure of information increments with the number of significant variables. Taking a gander at the ebb and flow malignant growth inquire about, one has a huge range of elements and data—and it is as yet expanding. Along these lines, it is inescapable to adapt to this heterogeneity and to construct huge examination bases by sharing and pooling therapeutic research information so as to acknowledge proof based customized drug. One approach to accomplish this objective could be the utilization of machine learning procedures.

Regression and Classification Techniques used for Cancer Diagnostics/Detection or Prediction: Ongoing advancements, for example, machine learning will act as cutting edge sequencing and had prepared for computational strategies and methods to assume complex and complicated jobs in such manner to define the narrative analogy for comprehensive approaches. Numerous essential issues in cell science require the thick nonlinear associations between useful modules to be considered. The significance of computer simulation to understand cell formation and broadly acknowledged cancer models and categories for assortment of reproduction of calculations helpful for concentrate certain subsystems have been structured for cancer detection and forecasting . In machine learning technique, information and yield is kept running on the machine simulation is meant to make a below mentioned generic framework depicted as figure No.1

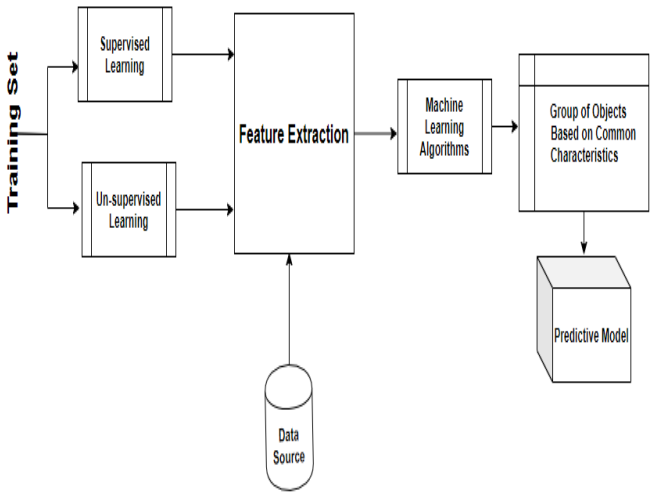

Figure: No.1 Schematics Representation of Machine Learning Models

II. RELATED STUDY

Over the previous decades, a constant development identified with disease inquire about has been performed [1]. Researchers connected distinctive strategies, for example, screening in beginning time, so as to discover sorts of malignancy before they cause side effects. In addition, they have grown new systems for the early forecast of malignancy treatment result. With the coming of new innovations in the field of medication, a lot of malignant growth information have been gathered and are accessible to the restorative research network. Be that as it may, the exact forecast of a sickness result is a standout amongst the most intriguing and testing errands for doctors. Thus, ML strategies have turned into a famous instrument for restorative scientists. These strategies can find and distinguish examples and connections between them, from complex datasets, while they can adequately foresee future results of a malignant growth type. Given the centrality of customized medication and the developing pattern on the utilization of ML procedures, we here present a survey of concentrates that make utilization of these techniques in regards to malignancy forecast and guess. In these investigations prognostic and prescient highlights are viewed as which might be autonomous of a specific treatment or are incorporated so as to direct treatment for malignant growth patients, separately [2]. Likewise, we talk about the sorts of ML techniques being utilized, the kinds of information they coordinate, the general execution of each proposed plan while we additionally examine their advantages and disadvantages. An undeniable pattern in the proposed works incorporates the coordination of blended information, for example, clinical and genomics. Nonetheless, a typical issue that we saw in a few works is the absence of outer approval or testing in regards to the prescient execution of their models. Unmistakably the use of ML strategies could improve the precision of malignant growth powerlessness, repeat and survival forecast. In view of [3], the precision of malignancy forecast result has essentially improved by 15%– 20% the most recent years, with the utilization of ML strategies. A few investigations have been

accounted for in the writing and depend on various techniques that could empower the early malignant growth finding and anticipation [4– 7]. In particular, these examinations depict approaches identified with the profiling of circling miRNAs that have been demonstrated a promising class for malignancy discovery and recognizable proof. Be that as it may, these strategies experience the ill effects of low affectability with respect to their utilization in screening at beginning periods and their trouble to segregate kindhearted from dangerous tumors. Different perspectives with respect to the expectation of disease result dependent on quality articulation marks are talked about in [8,9]. These investigations list the potential just as the constraints of microarrays for the forecast of malignant growth result. Despite the fact that quality marks could fundamentally improve our capacity for guess in malignant growth patients, poor advancement has been made for their application in the facilities. Be that as it may, before quality articulation profiling can be utilized in clinical practice, contemplates with bigger information tests and progressively satisfactory approval are required. In the present work just examinations that utilized ML methods for displaying malignancy analysis and forecast are exhibited. 2. ML strategies ML, a part of Artificial Intelligence, relates the issue of gaining from information tests to the general idea of surmising [10– 12]. Each learning procedure comprises of two stages: (I) estimation of obscure conditions in a framework from a given dataset and (ii) utilization of evaluated conditions to foresee new yields of the framework. ML has likewise been demonstrated a fascinating region with regards to biomedical research with numerous applications, where an adequate speculation is gotten via looking through a n-dimensional space for a given arrangement of organic examples, utilizing diverse methods and calculations [13]. There are two principle regular kinds of ML strategies known as (I) regulated learning and (ii) unsupervised learning. In regulated learning, a named set of preparing information is utilized to gauge or guide the info information to the ideal yield. Interestingly, under the unsupervised learning techniques, no marked precedents are given and there is no idea of the yield amid the learning procedure. Subsequently, it is up to the learning plan/model to discover designs or find the gatherings of the info information. In directed learning this system can be thought as an order issue. The undertaking of arrangement alludes to a learning procedure that sorts the information into a lot of limited classes. Two other normal ML errands are relapse and bunching. On account of relapse issues, a learning capacity maps the information into a genuine esteemed variable. Hence, for each new example the estimation of a prescient variable can be assessed, in light of this procedure. Bunching is a typical unsupervised assignment in which one attempts to discover the classes or groups so as to portray the information things. In view of this procedure each new example can be relegated to one of the distinguished groups concerning the comparable qualities that they share. Assume for instance that we have gathered therapeutic records applicable to bosom disease and we endeavor to foresee if a tumor is dangerous or considerate dependent on its size. The ML question would be alluded to the estimation of the likelihood that the tumor is harmful or no (1 = Yes, 0 = No). Delineates the order procedure of a tumor is threatening or not. The orbited records portray any misclassification of the kind of a tumor created by the system. Another kind of ML strategies that have been generally connected is semi-regulated realizing, which is a mix of directed and unsupervised learning. It consolidates marked and unlabeled information so as to develop a precise learning model. Normally, this sort of learning is utilized when there are more unlabeled datasets than named. While applying a ML technique, information tests establish the fundamental parts. Each example is portrayed with a few highlights and each component comprises of various kinds of qualities. Besides, knowing ahead of time the particular sort of information being utilized permits the correct choice of instruments and procedures that can be utilized for their examination. A few information related issues allude to the nature of the information and the preprocessing ventures to make them progressively reasonable for ML. Information quality issues incorporate the nearness of clamor, anomalies, absent or copy information and information that is one-sided unrepresentative. While improving the information quality, ordinarily the nature of the subsequent examination is likewise improved. Moreover, so as to make the crude information progressively appropriate for further investigation, preprocessing steps ought to be connected that emphasis on the change of the information. Various diverse procedures and techniques exist, pertinent to information preprocessing that emphasis on changing the information for better fitting in a particular ML strategy. Among these methods, probably the most imperative methodologies incorporate (I) dimensionality decrease (ii) include determination and (iii) highlight extraction. There are numerous advantages with respect to the dimensionality decrease when the datasets have countless. ML calculations work better when the dimensionality is lower [14]. Moreover, the decrease of dimensionality can dispense with immaterial highlights, lessen commotion and can deliver increasingly vigorous learning models because of the inclusion of less highlights. When all is said in done, the dimensionality decrease by choosing new highlights which are a subset of the old ones is known as highlight choice. Three primary methodologies exist for highlight determination in particular installed, channel and wrapper approaches [14]. On account of highlight extraction, another arrangement of highlights can be made from the underlying set that catches all the critical data in a dataset. The production of new arrangements of highlights takes into consideration gathering the portrayed advantages of dimensionality decrease.

III. PROPOSED SYSTEM

The clinical information taken from UCI (University of California, Irvine) entryway is utilized for further handling to get the ideal result. The proposed procedure worked here containing four stages. At first, pre-processing of information or data repository, second feature extraction and forming data relation, data regression using Linear Regression and lastly, classification using KNN the results will be evaluated with existing data models for accuracy and efficiency there in after

Stages

The clinical information taken from UCI (University of California, Irvine) entryway is utilized for further handling to get the ideal result. The proposed procedure worked here

containing four stages. At first, pre-processing of information or data repository, second feature extraction and forming data relation, data regression using Linear Regression and lastly, classification using KNN the results will be evaluated with existing data models for accuracy and efficiency there in after.

Residual Values

The residual values known to be errors are produced during regression. These values are the difference between the observed values and predicting values. The residual values can be 0, if it passes through the graph. The positive residual values shows that actual is more than predicted one. The negative residual values show that you have over predicted than the actual value.

Sum Of Squares

It is calculated by finding the sum of squared differences. Small sum of square indicates better models because there is less variation in the data.

Confusion Matrix

The confusion matrix also known as contingency table or error matrix calculates the precision and the performance of a classifier model on the given data set for those whose actual values are known. It shows four outcomes such as false negative, false positive, true negative and true positive.

Train Error

The train error is the error that you will get when you run the trained model back on the training data set and used in estimating model parameters. The large data is used for training.

Test Error

The test error is the error that you will get when you run the trained model on the data set that is not been exposed. This error measure the accuracy of the given model before shipping to the production. The small proportion of data is used for testing.

Existing Parameters

Patient Id: Foundation of competent healthcare, matching of patient to an predetermined treatment. Risk may occur if patient id is mismatched with the intended person.

Radius: The area where may cancer is diagnosed adhered as radius

Texture: Texture is a signs of breast cancer such as change in skin texture like puckering or dimpling redness or rash on the skin.

Perimeter: Boundary that defines the scope of particular process or activity.

Area: The surface where cell begins to grow out of control causing tumor that can be felt as a lump or seen by an x-ray.

Smoothness: The portion where additional organ is elevated and gathered in form of gilt.

Compactness: It is defined as the ratio of the volume and surface area.

Concave Point: Severity of concave portions of the contour.

Symmetry: It is a paired traits such as breast volume. Perfect symmetry can be disturbed by number of factors such as secretion of hormones.

Fractal Dimension: It is the ratio providing a statistical index of complexity comparing how the details of the pattern are changing with the scale at which it is measured.

IV. IMPLEMENTATION

Environment

The proposed scheme is implemented using Machine Learning Techniques such as Linear Regression and Support Vector Machine using Java Language on Eclipse integrated development environment(IDE). The Database(RDBMS) used are MySql/Sql Server/Sqlite. The Server used is Apache Tomcat. Script-lets are written Java Server Pages using Java Beans

V. COMPARISON AND ANALYSIS

Ruolan Xu, and Qiongjia Xu of stanford [15] evaluated constant advancement with cancer disease dataset is gotten from UCI database and gathered from Wisconsin medical clinic. There are 569 records altogether, with 212 harmful cases and 357 benevolent cases. Each column contains 30 distinct highlights and the determination of bosom malignant growth (0 for considerate and 1 for threatening). The 30 highlights speak to the mean, standard deviation and the most exceedingly terrible of 10 diverse cytopathology estimations, including, sweep, surface, edge, region, smoothness, conservativeness, concavity, sunken focuses, symmetry, and fractal measurement. Because of the little size of the dataset, we just have a preparation set and test set. 569 perceptions are part to 70% for preparing and 30% using certain machine learning algorithm therefore the train error and test error percentage is depicted in table no.1 for ready reference.

Train and Test Error for Different Models		
Model	**Train Error %**	**Test Error %**
Logistic Regression	3.6	5.1
LinearSVC	9.4	10.2
Random Forest Classifier	0.2	5.2
Naive Bayes	5.5	6.2
SVM with linear kernel	2.7	4.9

Table No.1 Train and Test Error for Different Machine Learning Models

Shockingly, LinearSVC performs very awful, while Naive Bayes performs moderately great. As indicated by sklearn

documentation, despite the fact that LinearSVC and SVM with straight bit both have a place with the SVM family, they are actualized utilizing two distinct libraries. Subsequently, LinearSVC is increasingly appropriate for bigger dataset with littler list of capabilities, while SVM with direct part works better for littler informational index (the time multifaceted nature is high however). For our situation, on the grounds that the dataset measure is impressively little and the list of capabilities is huge, LinearSVC doesn't accommodate our framework well, notwithstanding for the preparation set. Essentially, the out of the blue great execution Naive Bayes could be because of the extent of our dataset. By and large, we don't anticipate that Naive Bayes should perform well when there are solid relationships between's the highlights. As per our warmth map, numerous highlights are unequivocally related. In this manner, we are shocked to see the great mistake rate for Naive Bayes.

VI. COMPARISON AND ANALYSIS

Data : Likewise above getting and extracting the knowledge base repository from UCI comprising the Cancer Patients data patients repository of 5 major category of cancer is elevated using attribute extraction i.e. Category of Cancer, Gender, Age, Obstruct, Performance, Adherence, Extent, Status and Nodes Consequently, using Linear Regression approach to find out the Test Error model based on predictive attribute the results are depicted below as confusion matrix:-

Model	Train Error %
Logistic Regression	3.6
LinearSVC	9.4
Random Forest Classifier	0.2
Naive Bayes	5.5
SVM with linear kernel	2.7
Linear regression	-3.50

Table No.2 Confusion Matrix of Training Error % espionaged through Linear Regression

The above confusion matrix is in evolvement to produce the better and effective training error percentage then the result produced by Ruolan Xu and Qiongjia Xu [15] as depicted in table no.3. Consequently in future using KNN on age range vide dataset the better and effective comparison will be produced which will be validated using precision and recall model. Liner regression has minimum train error value as compared to other training models of machine learning.

REFERENCES

[1] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144: 786 Q2 646–74.

[2] Polley M-YC, Freidlin B, Korn EL, Conley BA, Abrams JS, et al. Statistical and practical considerations for clinical evaluation of predictive biomarkers. J Natl Cancer Inst 2013;105:1677–83.

[3] Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Informat 2006;2:59.

[4] Fortunato O, Boeri M, Verri C, Conte D, Mensah M, et al. Assessment of circulating microRNAs in plasma of lung cancer patients. Molecules 2014;19:3038–54.

[5] Heneghan HM, Miller N, Kerin MJ. MiRNAs as biomarkers and therapeutic targets in cancer. Curr Opin Pharmacol 2010;10:543–50.

[6] Madhavan D, Cuk K, Burwinkel B, Yang R. Cancer diagnosis and prognosis decoded by blood-based circulating microRNA signatures. Front Genet 2013;

[7] Zen K, Zhang CY. Circulating microRNAs: a novel class of biomarkers to diagnose and monitor human cancers. Med Res Rev 2012;32:326–48.

[8] Koscielny S. Why most gene expression signatures of tumors have not been useful in the clinic. Sci Transl Med 2010;2 [14 ps12-14 ps12].

[9] Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 2005;365:488–92.

[10] Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.

[11] Mitchell TM. The discipline of machine learning: Carnegie Mellon University. Carnegie Mellon University, School of Computer Science, Machine Learning Department; 2006.

[12] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. Morgan Kaufmann; 2005.

[13] Niknejad A, Petrovic D. Introduction to computational intelligence techniques and areas of their applications in medicine. Med Appl Artif Intell 2013;51.

[14] Pang-Ning T, Steinbach M, Kumar V. Introduction to data mining; 2006

[15] Ruolan Xu and Qiongjia, Xu Applying Different Machine Learning Models to Predict Breast Cancer Risk, https://www.semanticscholar.org/author/Ruolan-Xu/35696368